UNIVERSITY OF ILLINOIS

GRADUATE COLLEGE

DIGITAL COMPUTER LABORATORY

REPORT NO. 113

REMARKS ON ERRORS IN FIRST ORDER ITERATIVE PROCESSES WITH FLOATING-POINT COMPUTERS

by

J. Descloux

March 22, 1962

This work was supported in part by the
National Science Foundation
under grant G16489

REMARKS ON ERRORS IN FIRST ORDER ITERATIVE PROCESSES WITH FLOATING-POINT COMPUTERS

We consider the iterative process given by

$$x_{n+1} = x_n + G(x_n) \tag{1}$$

with limit r. All quantities are scalar. We suppose the convergence linear, i.e. there exists $0 \leq b < 1$ such that

$$| x + G(x) - r | \leq b | x - r | \quad \text{for every } x \tag{2}$$

Although analogous results can be probably obtained for other types of floating-point arithmetic, we suppose we are dealing with a binary computer with following properties:

1. All numbers, but 0, are of the form $\alpha\, 2^{\beta}$, where α is an exact binary fraction of N bits and the sign, β is an integer and $0.5 \leq \alpha < 1$.
2. There is a real zero represented for example by $\alpha = 0$, $\beta = -P$, where $-P$ is the smallest value of β; consequently the smallest non-zero numbers in absolute value are $\pm\, 2^{-P-1}$.

All these numbers, including zero, will be called "<u>normalized</u>".

Suppose that (1) is realized on this computer under the following assumptions:

1. The value effectively computed instead of $G(x)$ is $\overline{G}(x)$ with
$$\overline{G}(x) = (1 + y)\, G(x) + \zeta \quad | \eta | \leq d \quad | \zeta | \leq a \tag{3}$$
ζ and y depend on x; y and ζ are independent of x.
2. $\overline{G}(x)$ and the successive approximations are always represented on the computer as normalized numbers.

The effective process can be written

$$Y_{n+1} = [\, Y_n + \overline{G}(Y_n)\,]_R \tag{4}$$

indeed by using multiple precision $Y_n + \overline{G}(Y_n)$ can be represented exactly, since it is a multiple of 2^{-P-1}; however, by assumption 2, the mantissa of Y_{n+1} has no more than N digits and $Y_n + \overline{G}(Y_n)$ must be rounded as indicated by $[\ ]_R$.

We concentrate on attention on the rounding procedure in (4).
We consider two types of rounding procedures:

1. <u>Normal rounding</u>: $Y_{n+1} = [Y_n + \bar{G}(Y_n)]_N$; Y_{n+1} is a normalized number such that $|Y_{n+1} - (Y_n + \bar{G}(Y_n))| = \min$;

   When two different normalized numbers satisfy the above relation, any of them can be chosen as Y_{n+1}.

2. <u>Anormalous rounding</u>: $Y_{n+1} = [Y_n + \bar{G}(Y_n)]_A$;
   if $\bar{G}(Y_n) \geq 0$ let

   Z be the smallest normalized number such that $Z \geq Y_n + \bar{G}(Y_n)$

   W be the greatest normalized number such that $W \leq Y_n + \bar{G}(Y_n)$;
   If $\bar{G}(Y_n) \leq 0$ let

   Z be the greatest normalized number such that $Z \leq Y_n + \bar{G}(Y_n)$

   W be the smallest normalized number such that $W \geq Y_n + \bar{G}(Y_n)$

   then $[Y_n + \bar{G}(Y_n)]_A = W$ if $W \neq Y_n$

   $[Y_n + \bar{G}(Y_n)]_A = Z$ if $W = Y_n$

   The following relations are rather evident:

1. $| Y_{n+1} - [Y_n + \bar{G}(Y_n)]_N | \leq 2^{-N} (Y_n + \bar{G}(Y_n))$ (5)
2. $| Y_{n+1} - [Y_n + \bar{G}(Y_n)]_A | \leq 2^{-N+1} (Y_n + \bar{G}(Y_n))$ (6)
3. if $Y_n < Y_n + \bar{G}(Y_n) < p$, then $Y_n < [Y_n + \bar{G}(Y_n)]_A < p$ (7)

   if $p < Y_n + \bar{G}(Y_n) < Y_n$, then $p < [Y_n + \bar{G}(Y_n)]_A < Y_n$ (8)

   where p is any number and provided there is a normalized number s such that $Y_n < s < p$ for (7) and $p < s < Y_n$ for (8).

Theorem a) By using the normal rounding for any Y_o, there exists a finite number M such that $| Y_n - r | \leq B_N =$

$$\frac{2^{-N} | r | + a (1 + 2^{-N})}{2 + 2^{-N} - (1 + d) (1 + b) (1 + 2^{-N})} \leq \frac{2^{-N} | r | + a}{1 - b - 2d - 2^{-N}}$$

for $n > M$.

b) By using the anormalous rounding, for any Y_o, there exists a finite number M such that

$$| Y_n - r | < B_A = (2^{-N+1} \; | r | + 2^{-P-1} + \frac{a\,(1 + 2^{-N+1})}{2 - (1 + d)\,(1 + b)} \,)$$

for $n > M$.

In both cases, if the bounds B_N or B_A are non-positive, they must be replaced by $+ \infty$.

<u>Truncation errors</u> Suppose we compute with infinite precision, i.e. without rounding errors.

The remaining inaccuracy of the process will be called the <u>truncation error</u> and comes from the errors η and ζ in equation 3.

We consider the limits of B_N and B_A when $N \to \infty$ and $P \to \infty$

$$B = \lim_{N \to \infty} B_N = \lim_{\substack{N \to \infty \\ P \to \infty}} B_A = \frac{a}{2 - (1 + d)\,(1 + b)}$$

Using analog agruments to these in the proof of the theorem, one can find the following result:

Let for any V_o, the sequence V_n be defined by

$$V_{n+1} = V_n + \overline{G}(V_n)$$

then any point of accumulation V of the sequence satisfy the relation $| V - r | \leq B$. We give an example where the bound is reached; let

$$G(x) = - (1 + b)\,(x - r)$$

$$G(x) = - (1 + d)\,(1 + b)\,(x - r) - a \frac{x - r}{| x - r |}$$

First we remark that if $a = 0$, the sequence will converge if and only if $| 1 - (1 + d)\,(1 + b) | < 1$, i.e. $2 - (1 + d)\,(1 + b) > 0$, since d and b are non-negative numbers; if the condition is not satisfied, the sequence diverges to infinity.

For $a \neq 0$, it is easy to verify that if

$$V_o = r + \frac{a}{2 - (1 + d)\,(1 + b)}$$

$$\text{then } V_1 = r - \frac{a}{2 - (1+d)(1+b)}$$

$$V_2 = r + \frac{a}{2 - (1+d)(1+b)}$$

In order to compare the results of the theorem, i.e. to compare B_A and B_N, first suppose that $a = 0$; then

$$B_A = |\, r \,|\, 2^{-N+1} + 2^{-P-1} \quad ; \quad B_N = \frac{|\, r \,|\, 2^{-N}}{2 + 2^{-N} - (1+d)(1+b)(1+2^{-N})} \quad ;$$

Since $d \geq 0$, $b \geq 0$ it follows $B_N \geq 1/2\, B_A - 2^{-P-1}$; B_A is independent of d and b and remain very small; for $d \cong 0$, $b \cong 0$, B_N is slightly smaller than B_A, but if $b \cong 1$, i.e. when the convergence is very slow, B_N can become very large. For reasonable values of b and d, the increase of value of the bounds B_A and B_N due to $a \neq 0$ are almost equal (i.e. if one neglects the effects of the rounding, i.e. if $N \to \infty$). Consequently the anomalous rounding can be considered as safer than the normal rounding.

<u>Remark</u> In the theorem a, it is asserted that $B_N \leq \dfrac{2^{-N}\,|\, r \,|\, + a}{1 - b - 2d - 2^{-N}}$. This "bound" of the bound B_N is useful when $b \cong 1$.

<u>Example</u> The bounds B_A or B_N can be reached only in trivial cases. However, for the general case, they remain realistic; that is true for B_A since B_A is not much greater than the truncation error; as for B_N, let us consider the following example:

Let $b = 3/4$, $d = 1/8$, $a = 5 \cdot 2^{-35}$, $N = 32$; $r = 3/4$

$$G(x) = -\,7/4\,(x - 3/4)$$

$$\bar{G}(x) = -\,9/8 \cdot 7/4\,(x - 3/4) - 5 \cdot 2^{-35} \cdot \text{sign}\,(x - 3/4)$$

$$B_N = 44 \cdot 2^{-32};\ B_A = 21.5 \cdot 2^{-32};\ B = 20.2^{-32},$$

it is easy to check the following computations:

$$Y_o = 3/4 \pm 32 \cdot 2^{-32}$$

$$Y_1 = [Y_o + \bar{G}(Y_o)]_N = 3/4 - 32 \cdot 2^{-32}$$

$$Y_2 = [Y_1 + \bar{G}(Y_1)]_N = 3/4 + 32 \cdot 2^{-32}$$

Proof of the Theorem

Lemma 1 Let W_1 and W_o satisfy the relation

$$W_1 = (1 + \epsilon)\,(W_o + (1 + \eta)\,(G(W_o) + \zeta))$$

where $|\zeta| < e =$ constant and η, $G(W)$, ζ satisfy the hypothesis given by the equation 2 and 3.

$$\text{Let } K = \frac{e\,|r| + a\,(1 + e)}{2 + e - (1 + d)\,(1 + b)\,(1 + e)}\;;$$

then : 1. if $|W_o - r| > K$, then $|W_1 - r| < |W_o - r|$

2. if $|W_o - r| \leq K$, then $|W_1 - r| \leq K$

Proof:

$$W_1 - r = (1 + \epsilon)\,(W_o + (1 + \eta)\,(G(W_o) + \zeta)) - r$$

$$= (1 + \epsilon)\,(1 + \eta)\,(W_o + G(W_o) - r) - \eta\,(1 + \epsilon)\,(W_o - r) + r\epsilon + \zeta\,(1 + \epsilon);$$

by equation 2:

$$|W_1 - r| \leq |(1 + \epsilon)\,(1 + \eta)|\,b\,|W_o - r| + |\eta(1 + \epsilon)|\,|W_o - r| + |r\epsilon| + |\zeta\,(1 + \epsilon)|$$

$$|W_1 - r| \leq |W_o - r|\,\{(1 + e)\,(1 + d)\,(1 + b) - 1 - e\} + |r|\,e + a(1 + c) \qquad (9)$$

First suppose $|W_1 - r| > K$; by 4:

$$|W_1 - r| \leq |W_o - r| - (2 + e - (1 + e)\,(1 + d)\,(1 + b))\,|W_o - r| + |r|\,e + a\,(1 + e)$$

$$< |W_o - r| - (2 + e - (1 + c)\,(1 + d)\,(1 + b))K + |r|\,e + a\,(1 + c) \leq |W_o - r| \qquad \text{q.e.d.}$$

Now suppose $|W_1 - r| \leq K$; by 4:

$$|W_1 - r| \leq K\,\{(1 + e)\,(1 + d)\,(1 + b) - 1 - e\} + |r|\,e + a\,(1 + c)$$

$$\leq K - K\,(2 + e - (1 + e)\,(1 + d)\,(1 + b)) + |r|\,e + a\,(1 + c) \leq K \qquad \text{q.e.d.}$$

Lemma 2 $B_N \leq \frac{2^{-N} | r | + a}{1 - b - 2d - 2^{-N}}$ (if the denominator ≤ 0, the expression must be replaced by $+ \infty$).

Proof

$$\frac{2^{-N} | r | + a}{1 - b - 2d - 2^{-N}} = \frac{(2^{-N} | r | + a) (1 + 2^{-N})}{(1 - b - 2d - 2^{N}) (1 + 2^{-N})}$$

$$= \frac{(2^{-N} | r | + a) (1 + 2^{-N})}{2 + 2^{-N} - (1 + b) (1 + d) (1 + 2^{-N}) - d (1 - b) (1 + 2^{-N}) - 2^{-2N}}$$

$$\geq \frac{2^{-N} | r | + (1 + 2^{-N}) a}{2 + 2^{-N} - (1 + b) (1 + d) (1 + 2^{-N})} = B_N \quad \text{q . e.d.}$$

Proof of theorem a By equation 5:

$$Y_{n+1} = (1 + \epsilon) (Y_n + (1 + \eta) G(Y_n) + \zeta) \text{ with } | \epsilon | \leq 2^{-N}.$$

by replacing in lemma 1 e by 2^{-N}, we find $K = \frac{2^{-N} | r | + a (1 + 2^{-N})}{2 + 2^{-N} - (1 + d) (1 + b) (1 + 2^{-N})}$

then the theorem a and the lemma 1 are equivalent, since there exists only a finite number of normalized numbers.

The lemma 2 completes the proof.

Proof of theorem b Since there exists only a finite number of normalized numbers, the theorem b is equivalent to the following assertions:

I. If $| Y_o - r | \leq \frac{a}{2 - (1 + d) (1 + b)}$, then $| Y_1 - r | < B_A$

II. If $\frac{a}{2 - (1 + d) (1 + b)} < | Y_o - r | < B_A$, then $| Y_1 - r | < B_A$

III. If $| Y_o - r | \geq B_A$, then $| Y_1 - r | < | Y_o - r |$

I. By lemma 1, $r - \frac{a}{2 - (1 + d) (1 + b)} \leq Y_o + \overline{G}(Y_o) \leq r + \frac{a}{2 - (1 + d) (1 + b)}$

since $2^{-N+1} (Y_o + \overline{G}(Y_o)) \leq (| r | + \frac{a}{2 - 1 (1 + d) (1 + b)}) 2^{-N+1}$, by

equation 6, we have:

$$r - \frac{a}{2 - (1 + d)(1 + b)} - \left(| r | + \frac{a}{2 - (1 + d)(1 + b)} \right) 2^{-N+1} \leq Y_1$$

$$\leq r + \frac{a}{2 - (1 + d)(1 + b)} + \left(| r | + \frac{a}{2 - (1 + d)(1 + b)} \right) 2^{-N+1}$$

and consequently $| Y_1 - r | \leq | r | \, 2^{-N+1} + \dfrac{a}{2 - (1 + d)(1 + b)} < B_A$

II. Suppose that $r + \dfrac{a}{2 - (1 + d)(1 + b)} < Y_o < r + B_A$ (the proof is analogous when $r - B_A < Y_o < r - \dfrac{a}{2 - (1 + d)(1 + b)}$). By lemma 1

$$2r - Y_o < Y_o + \overline{G}(Y_o) < Y_o,$$

$$r - B_A < Y_o + \overline{G}(Y_o) < Y_o;$$

but $r - B_A \leq r - | r | \, 2^{-N+1} - 2^{-P-1} < r < Y_o$ and there exists a normalized number s such that $r - | r | \, 2^{-N+1} - 2^{-P-1} < s \leq r$; we apply equation 8:

$$r - B_A < [Y_o + \overline{G}(Y_o)]_A < Y_o < r + B_A, \text{ i.e. } | Y_1 - r | < B_A \quad \text{q.e.d.}$$

III. Suppose $Y_o > r + B_A$ (the proof is analogous when $Y \leq r - B_A$). By lemma 1:

$$2\,r - Y_o < Y_o + \overline{G}(Y_o) < Y_o;$$

but $2r - Y_o \leq r - 2^{-N+1} | r | - 2^{-P-1} < r < Y_o$ and there exists a normalized number s such that $r - | r | \, 2^{-N+1} - 2^{-P-1} < s \leq r$; we apply equation 8:

$$2r - Y_o < [Y_o + \overline{G}(Y_o)]_A < Y_o, \text{ i.e. } | Y_1 - r | < | Y_o - r | \quad \text{q.e.d.}$$